Are we biased on bias?
Characterizing social bias research in the ACL community

Anonymous EMNLP submission

Abstract

Recent events in business, politics and society have shed light on the importance and potential dangers of Natural Language processing (NLP) in the real world. NLP applications have gained unprecedented popularity not just among scientist and practitioners, but also the general public. As we develop new methodologies and curate new benchmarks and datasets it is more important than ever to consider the implications and societal impact of our work. In this paper, we characterize the landscape of societal bias research within the ACL community and provide a quantitative and qualitative survey by analyzing an categorized corpus of *348* papers. More specifically, we present a definition of social bias based on ethical principals and investigate (i) types of bias, (ii) languages, and (iii) type of paper. We find that there is significantly more work on gender biases and English than other languages. Finally, we discuss the possible causes behind our findings and provide pointers to future opportunities.

1 Introduction

Traditionally, the NLP community has focused on ethical debates around privacy (Hovy and Spruit, 2016) ensuring that data is anonymized appropriately. More recently, there has been increased awareness that NLP research has a direct impact on peoples' lives (Mayfield et al., 2019; Bender and Friedman, 2018). For example, summarization systems can amplify misinformation (Smiley et al., 2017), and sentiment analysis (SA) systems can assign more negative sentiments/emotions based on race and/or gender (Kiritchenko and Mohammad, 2018). While such research used to be more academic (Leidner and Plachouras, 2017), these concerns are having an increasing impact in industry (Schnoebelen, 2017; Jin et al., 2021b) with consequences for users (Prabhumoye et al., 2021). It is well known that language data encodes demographics and biases (Bender and Friedman, 2018). There is a risk that using such data can disclose inappropriate information about particular individuals, as well as undesirable attitudes towards individuals and groups (Hovy and Spruit, 2016; Eckert and Rickford, 2001) and social hierarchies (Blodgett et al., 2020). There are also concerns that systems based on inappropriate data are likely to repeat such biases, and may even amplify them (Bender and Friedman, 2018). In this paper, we survey **348** papers collected from the ACL anthology that focus on social bias and ethics in NLP research. We make three kinds of contributions, where (i) we present a working definition of social bias from a philosophy perspective, (ii) quantify our findings by annotating our corpus of papers and (iii) provide a discussion and pointers for possible future research directions. Through a quantitative analysis of current trends we attempt to answer the following questions:

- What kind of social biases is the ACL community concerned with?
- How many languages is bias studied in?
- What types of papers are present?

2 Related previous surveys

There is a considerable literature on social biases in NLP. Much of this work provides guidelines and/or recommendations. One of the first position papers on the topic outlined the need for ethical considerations that go beyond privacy concerns for users, and focus on the social impact of experiments and applications on individuals and (minority) groups Hovy and Spruit (2016).

Surveys on social bias emphasize a variety of different aspects, such as embedding representations, data collection and annotation, downstream task performance, metrics and limiting negative impact (Garrido-Muñoz et al., 2021; Mohammad, 2022b; Schnoebelen, 2017). Work by Bender and Friedman (2018); Hovy and Prabhumoye (2021) outline the concept of data statements and sources of bias respectively to aid the research design process. Other research has reviewed how to mitigate bias (Chandrabose et al., 2021; Meade et al., 2022; Balkir et al., 2022), how to teach bias, ethics and privacy to students (Bender et al., 2020; Friedrich and Zesch, 2021), evaluate existing metrics (Czarnowska et al., 2021; Delobelle et al., 2022), handle challenges presented by new laws (e.g., GDPR) (Lewis et al., 2017) and apply existing principles from ethics and privacy to NLP (De Jong et al., 2018; Leidner and Plachouras, 2017; Prabhumoye et al., 2021). Our work follows the precedent established by Blodgett et al. (2020), who used keywords to select papers from the ACL anthology, and then enlarged the sample by following citations to other popular venues (eg AAAI, ICML etc.). Similarly, we also align with Field et al. (2021) who solely focus on papers published at ACL venues but draw on conclusions from NLP papers published elsewhere. Relying on keywords, of course, introduces possibilities for false positives and false negatives.

Scope of this survey In this work, we solely focus on papers published in the ACL anthology to gain a better insight into current trends and popular research approaches for social bias. One limitation of such an approach is that seminal work published at other venues is not reviewed here. In Appendix 7, we provide a table that shows each paper reviewed in this survey.

3 Defining Bias

The dictionary definition of bias is "an inclination or prejudice for or against one person or group, especially in a way considered to be unfair" (Stevenson, 2010). Based on this, defining bias involves an interaction among different components: (i) individuals or categories that determine a group, (ii) attitudes towards this group, and (iii) assessment of this attitude in relation to fairness. Bias comes into existence when a specific attitude is formed, which may or may not be fair. For example, in investigating a disease that is more prevalent among women, use of *gender* is a relevant variable and, in itself, does not entail a differential attitude. Once a differential attitude is formed however, bias comes into existence and the attitude may be negative or positive. By definition, if a distinctly positive attitude is formed about one variable (e.g., *gender*), it entails a less favorable attitude about the other *gender* categories. This does not mean that all bias is necessarily unfair, there are multiple theories and definitions of fairness that are formulated and analyzed in-depth in political philosophy (Lamont and Favor, 2004). The formal principle of equality formulated by Aristotle (Ameriks and Clarke, 2000) states that equals must be treated equally, which is often referred to as 'the fairness ideal', but it is neither a prevailing definition nor a useful one in practice. Without identifying relevant features, such a definition would not prevent categories such as *race* or *wealth* to be used as variables for differential treatment. However, a fairness approach (Lamont and Favor, 2004) based on equal opportunity might require a differential attitude (i.e., bias) towards a certain category in order to 'level the playing field'. For example, if women are routinely given worse performance reviews and lower pay for successfully completing the same tasks as men, then there is differential treatment. Meaning women who are as successful as men cannot have the same opportunities. According to this understanding of fairness, a bias towards women would be fair. It is also worth noting that positive or negative discrimination is distinct from bias. While discrimination is about treatment, bias is about attitude. In other words, bias may lead to discrimination (Mateo and Williams, 2020). In this context, the ethical issue about bias is tied to differential treatment of descriptive categories resulting in unfair outcomes. Identifying and dealing with ethical concerns related to bias, must necessarily involve identifying the descriptive categories and the biases against those categories as well as examining whether the said bias is unfair, according to the relevant definition of fairness.

4 Methodology

Following the standard practice mentioned above, we searched the ACL anthology[1] in September 2022 for relevant titles and abstracts using the keywords: *ethical, ethics, fairness, fair, bias, social, society, societal, social good*. For papers, published after September 2022 we manually screened all conference proceedings for the same keywords. One limitation of relying on a keyword search alone is that we might miss any work that refers to a bias directly in the title, for example 'fatphobia detection in online forums'.

[1] https://github.com/shauryr/ACL-anthology-corpus

Filtering Strategy A total of *1,437* papers were returned by the search; *523* papers were retained after a manual screening of titles and abstracts. We removed duplicates, as well as work not related to bias and/or ethics in NLP. Then we downloaded full papers, and filtered out papers if: (i) the contribution was a talk, demonstration, abstract or keynote, (ii) "bias" was used in the machine learning sense, or (iii) the paper did not focus on social bias. This process produced a collection of *348* papers.

Categorization process We identify trends in our corpus by empirically determining a set of five categories, where we fully review each paper manually. We make our corpus alongside with its corresponding categories and labels publicly available[2]. We focus on four elements for each paper to identify trends, where we identify language investigated, type of bias analyzed and what kind of paper is introduced and which NLP area it belong to. For the type of paper analysis we utilize the authors description of contributions to split the papers in the following categories:

- **Method:** In this category of papers, the main focus of the work is to contribute a new method, which includes but is not limited to novel ways to measure or mitigate social bias.

- **Analysis** Papers in this category, examine existing datasets, benchmarks, language models, NLP systems or embeddings for bias using social science methods, statistics and mixed methods approaches. For example, authors who have conducted research in this category have explicitly stated that they conduct an analysis or outline a mixed method approach.

- **Surveys and Position papers:** This category of papers includes surveys, guides, tutorials, reviews and position papers.

- **Dataset, Benchmarks or Resources:** Paper in this category propose new datasets, benchmarks, lexicons, challenge sets and often include some preliminary analysis of the new data either collected through crowd-sourcing.

- **Datasets, Benchmarks and Methods:** This is a combination of papers that focus on both introducing a new resource (e.g.: dataset or benchmark) in addition to a new methodology.

[2]Link-added-upon-publication

Year	Papers	Year	Papers
2010-2016	11	2020	68
2017	16	2021	96
2018	10	2022	79
2019	46	2023	22
Totals	83	**Totals**	265

Table 1: 76,15% of the 348 papers are from 2020-2023.

5 Empirical Findings

The 348 papers were published between 2010 and 2023. Table 1 shows that there has been considerable growth in interest in bias research.

Type of Paper Based on the criteria outlined above, we have found that the majority of papers introduced are *Method* papers (57.75%), followed by *Datasets and Benchmarks* (20.40%), *Surveys* (15.22%), *Analysis* (4.31%) and combined *Datasets, Benchmarks and Methods* (2.29%). We take a closer look the majority category *Methods*, where we assigned *Method* papers into different categories based on the contribution of the paper (see Table 2).

Methods We distinguish between bias detection and measurement, as detection does not necessarily measure or remove any kind of bias. We found similarly to (Blodgett et al., 2020) that many papers propose a combination of techniques, hence why we have decided to merge such approaches into the category *'Measurement and Mitigation'*. Many works in Debiasing and Mitigation apply or extend methods such as WEAT (Caliskan et al., 2017) or HARD DB (Bolukbasi et al., 2016) to a specific benchmark or dataset. One negative side effect of this could be that for example in *gender bias* Measurement/Mitigation, there is no evidence that HardDB can or should be applied to languages with grammatical gender (Sun et al., 2019). There has also been criticism of removing bias (Caliskan et al., 2017), where removing bias (i) only changes how an application or algorithm understands the world but not how it applies the knowledge gained from understanding ('fairness through blindness'), (ii) could harm meaning and accuracy and (iii) bias can (unintentionally) be introduced through other avenues during the design process. Therefore, simply removing bias is not enough (Chandrabose et al., 2021) and developing new methods requires consistent reflection as bias in NLP systems is never fully inescapable (Waseem et al., 2021). Work by (Hovy, 2015; Kiritchenko and Mohammad, 2018) has found that the some information bias mitigation techniques can be beneficial in improving performance in downstream tasks. The methods in the *Miscellaneous* category take a different approach to working on bias. For example (Fisher et al., 2020) in including bias sensitive attributes by defining a whitelist of triples for uncontroversial cases. Arguably, one drawback here is that it is up to the person whitelisting to decide what is not controversial and what is. Similarly, (Touileb et al., 2021) and (Wang et al., 2017) use *gender* and *linguistic bias* respectively to improve classification results. Touileb et al. (2021) first show how female critics disproportionately give lower ratings to female authors, where removing metadata may have the opposite effect in that it does not help traditionally underrepresented groups in a specific domain. At the same time, many of the methods looking at measuring or removing bias complicate the data and tasks at hand and can lead to the development of systems that are not reliable when used in a more complex context (Talat et al., 2022b). This also applies to the predominant use of intrinsic metrics in bias measurement. These metrics may shed more light on how much bias exists in a dataset/LM, but does not necessarily correlate with performance on downstream tasks and therefore does not show the true harm of bias (Orgad and Belinkov, 2022). Thus, we may run the risk of developing methods for each new dataset or benchmark and missing out on crucial information that shows how bias affects different downstream tasks in different ways. However, documenting and measuring bias in a systematic way is crucial to understanding what harms can be caused in real life situations, so that preventive methods can be developed (Dev et al., 2021b). Current approaches in mitigation and/or measurement methods are evaluated on a variety of NLP areas, including Language Models (35.74%), Classification (22.85%) , NLG (14.76%) and NLI (3.33%). It is unsurprising that much attention has been paid to embedding representations that are trained on large amounts of text (Kiritchenko and Mohammad, 2018; Mayfield et al., 2019; Talat et al., 2022b). This has the benefit of bias methods being more widely applicable, but it also means that there are distinct limitations when a method is tied to a specific architecture rather than the task/benchmark itself. It means bias measures are no longer comparable in relation to other benchmarks and bias can be introduced at any stage of an NLP system design as it depends on where and how the final LLM is applied and to which community (Talat et al., 2022b). An important trade-off to consider is the balance between generalizable and context-sensitive methods to measure bias in downstream tasks. There are also other areas that have done work on bias but are not represented as well in this survey, which include but are not limited to Speech Recognition (Kwako et al., 2022; Savoldi et al., 2022), Multi-Modal NLP (Chen et al., 2020a; Srinivasan and Bisk, 2022a) and Information Extraction (Li et al., 2022b; Sun and Peng, 2021).

Type	Papers	%
Measurement	98	46.66
Debiasing / Mitigation	52	24.76
Combinations of above	32	15.23
Detection	19	9.04
Generation	5	2.38
Miscellaneous	4	1.9
Total	210	100.0

Table 2: Empirical taxonomy of methods.

Social biases In Figure 1 we show the types of biases investigated, where we include all social biases that occur more than once (see diagonal of the matri). Furthermore there are a small number of biases that only occur once and are not shown in the table, such as bias transfer hypothesis (Steed et al., 2022) or dialect bias (Tatman, 2017b). Furthermore, we included the category *multiple social biases*, where the paper does not explicitly list or describe the specific type of bias investigated (Ghosh et al., 2021; Ramponi and Tonelli, 2022; Mireshghallah and Berg-Kirkpatrick, 2021; Loukina et al., 2019). There also is the *intersectional bias* category, which shows how different elements of a person's identity (e.g., gender, race and age) can either be a benefit or disadvantage and lead to compounded discrimination (Crenshaw, 2013). For example, recent work by Lalor et al. (2022) benchmarks a variety of NLP models on different downstream tasks for its performance on intersectional biases and (Câmara et al., 2022) introduce a framework for unisectional and intersectional analysis of sentiment analysis in a multilingual setting. Few works focus on biases other than *gender*, where (Davidson et al., 2019; Sap et al., 2019; Manzini et al., 2019) look at *racial bias* and (Hutchinson and Mitchell, 2019; Herold et al., 2022) investigate *disability bias*. Noticeable is that some biases are not investigated on their own, such as *age*, *religion*, *sexuality* or *profession*. We included *political/media bias* in this analysis, if the paper also looks at social bias. For example, debiasing claims that include attitudes towards a group (e.g., *sexuality*). However, this type of work does not explicitly mention social biases when attitudes or characteristics of the targeted group are only implied. Dayanık and Padó (2020) looks media bias on a immigration dataset (MARDY) but does not mention implied social biases (e.g., *nationality* or *ethnicity*). The most frequently combined social biases are *gender* and *race*.The most frequently combined social biases are *gender* and *race*. In Table 3 we compute residuals between observed joints and predictions on margin, where highlighted in *green* are highly saturated areas and shades of *red* show less saturated areas.

Languages We show the languages used in each type of paper, excluding *Surveys and Position papers* in Figure 2. There are a total of 34 languages, however we leave out any languages that only occur once in the visualization. There are 11 languages not visualized, including *Farsi*, *Urdu*, *Wolof*, *Bengali*, *Armenian*, *Bengali*, *Inuktitut*, *Ukrainian*, *Hungarian*, *Indonesian* and *Lithuanian*. English, German, Spanish and Chinese are most commonly used either on their own or in combination with each other. The majority of all papers focus on a single language at a time. Furthermore, the vast majority of LLMs are monolingual and do not encode the cultural variety that naturally occurs within one language, for example non-standard English varieties (Talat et al., 2022b). Therefore it is important to not only document the type of bias investigated, but also contextualize bias within a language's cultural context, understanding of said bias and document the language itself (*Bender Rule* (Bender, 2011)). Based on this collection, bias research is heavily biased towards western and Anglo-centric notions of bias and very few works focus on non-English benchmarks (Talat et al., 2022b; Hovy and Spruit, 2016). This proves extremely problematic when English benchmarks are automatically translated, but many of the biases do not hold true in non-Western cultural contexts. For example, gendered professions do not necessarily translate across every language or culture (Talat et al., 2022b) and many NLP systems trained on written English (e.g., Penn Tree Bank) do not perform well on non-standard English (Mayfield et al., 2019). From Table 4 we can see the residuals between observed joints showing a clear over-saturation (*green*) for specific combinations of languages (e.g: *English* and *German* or *English* and *Spanish*).

6 Discussion

Datasets and Benchmarks Previous work (Hovy and Spruit, 2016; Hovy, 2018; Talat et al., 2022b) outlined a number of reasons that may explain why there are so many papers focusing on the same datasets, benchmarks, biases, and languages. In the following section, we highlight some of the elements that may explain why there is an uneven distribution of work and resources.

- **Experimental setup** The majority of of work in bias research has focused on using intrinsic bias measurements (bias in internal model representations) and little attention has been paid to extrinsic metrics (Orgad and Belinkov, 2022; Delobelle et al., 2022). A very real consequence of this is that much work does not appropriately describe, contextualize and identify the potential harms that bias has in real-world scenarios (Blodgett et al., 2020). Also bias is inadvertently introduced in intrinsic metrics, where lexicons used to measure bias in one dataset produces very different results in another (Antoniak and Mimno, 2021). Similarly, (Goldfarb-Tarrant et al., 2021) have found that there is no correlation between intrinsic and extrinsic metrics. Another key problem is that often newly proposed datasets are linked to specific metrics, which makes it hard to draw conclusions from individual case studies as many results are not generalizable (Orgad and Belinkov, 2022). Another element impacting metrics is the composition of test data, where (Orgad and Belinkov, 2022) found that test sets often don't contain balanced examples. However, most metrics are defined over a whole dataset and are therefore sensitive to its composition, which may lead to variability in results. Both metric choice and dataset composition can significantly change the results and conclusions drawn from a downstream task or dataset (Akyürek et al., 2022a). Therefore, it is important to (i) provide motivation for including / excluding a particular metric and describe how it impacts downstream performance, (ii) compare a metric across a variety of datasets and (iii) compare many metrics across individual datasets.

- **Funding** There are unintended consequences of research that can be traced to how research projects are funded (e.g., governments or mil-

gender	race	religion	age	sentiment	profession	media	nationality	political	multiple social	annotator	sexuality	disability	ethnicity	physical appearance	dialect	human	intersectional	
	19.3	8.2	8.8	-8.5	5.8	-8.6	3.0	-6.5	-7.5	-6.5	5.1	1.7	3.8	3.0	1.1	-2.3	-1.7	gender
19.3		12.2	12.4	-1.0	6.8	-3.0	4.2	-1.6	-2.6	-2.6	8.6	3.8	4.2	4.6	0.0	-0.8	-0.6	race
8.2	12.2		4.4	-1.4	7.9	-1.0	7.0	-0.9	-0.9	-0.9	4.2	4.2	2.4	5.5	0.7	-0.3	-0.2	religion
8.8	12.4	4.4		-1.3	3.9	-1.0	6.1	0.1	-0.9	0.1	4.2	5.3	4.4	5.5	0.7	-0.3	-0.2	age
-8.5	-1.0	-1.4	-1.3		0.1	0.1	0.2	-0.7	-0.7	-0.7	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	0.8	sentiment
5.8	6.8	7.9	3.9	0.1		-0.7	5.4	-0.6	-0.6	-0.6	3.4	4.5	0.6	4.7	-0.2	-0.2	-0.1	profession
-8.6	-3.0	-1.0	-1.0	0.1	-0.7		-0.6	-0.6	-0.6	-0.6	-0.5	-0.5	-0.4	-0.3	-0.2	-0.2	-0.1	media
3.0	4.2	7.0	6.1	0.2	5.4	-0.6		-0.5	-0.5	-0.5	3.5	5.6	3.6	4.7	-0.2	-0.2	-0.1	nationality
-6.5	-1.6	-0.9	0.1	-0.7	-0.6	-0.6	-0.5		-0.5	0.5	0.6	0.6	0.7	-0.3	-0.2	-0.1	-0.1	political
-7.5	-2.6	-0.9	-0.9	-0.7	-0.6	-0.6	-0.5	-0.5		-0.5	-0.4	-0.4	-0.3	-0.3	-0.2	-0.1	-0.1	multiplesocial
-6.5	-2.6	-0.9	0.1	-0.7	-0.6	-0.6	-0.5	0.5	-0.5		-0.4	-0.4	-0.3	-0.3	-0.2	-0.1	-0.1	annotator
5.1	8.6	4.2	4.2	-0.7	3.4	-0.5	3.5	0.6	-0.4	-0.4		4.6	0.7	3.8	-0.2	-0.1	-0.1	sexuality
1.7	3.8	4.2	5.3	-0.6	4.5	-0.5	5.6	0.6	-0.4	-0.4	4.6		1.7	4.8	-0.2	-0.1	-0.1	disability
3.8	4.2	2.4	4.4	-0.5	0.6	-0.4	3.6	0.7	-0.3	-0.3	0.7	1.7		0.8	-0.1	-0.1	-0.1	ethnicity
3.0	4.6	5.5	5.5	-0.4	4.7	-0.3	4.7	-0.3	-0.3	-0.3	3.8	4.8	0.8		-0.1	-0.1	-0.1	physicalappearance
1.1	0.0	0.7	0.7	-0.3	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.1	-0.1		-0.1	-0.0	dialect
-2.3	-0.8	-0.3	-0.3	-0.2	-0.2	-0.2	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1		-0.0	human
-1.7	-0.6	-0.2	-0.2	0.8	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.0	-0.0		intersectional

Table 3: Observed joints: number of papers with combinations of languages (ISO 639). biases

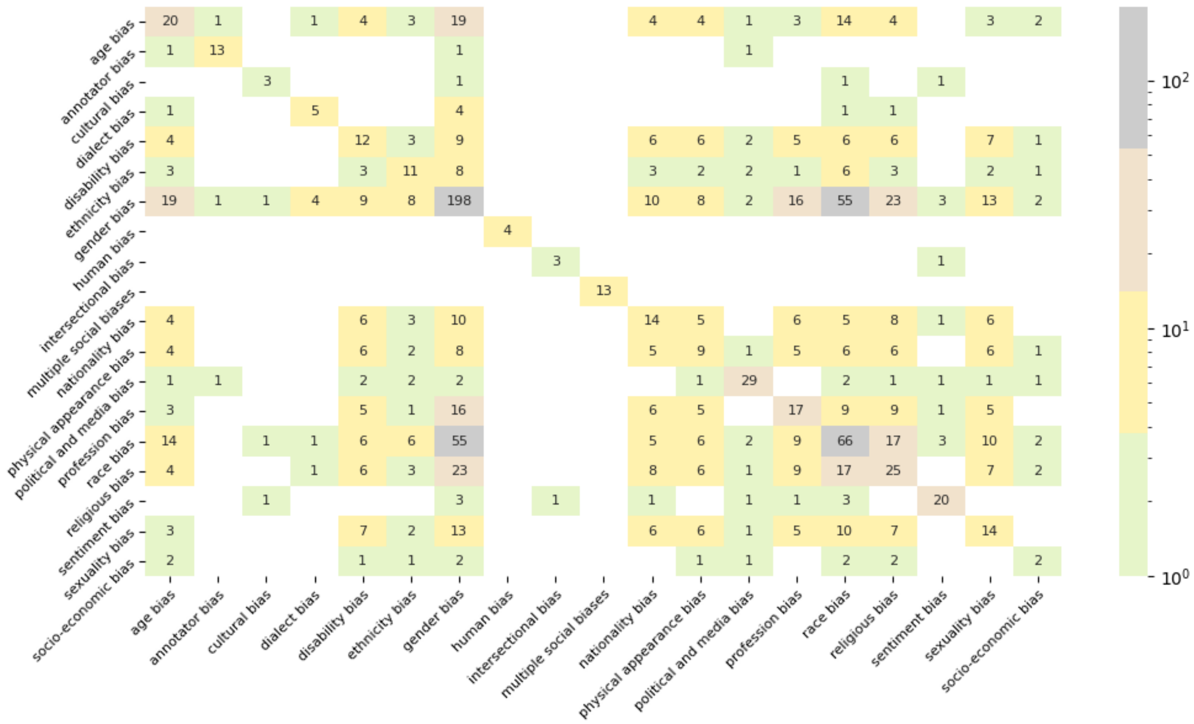

Figure 1: A log-scaled heatmap showing the type and frequency of social biases.

itary interests), where researchers should be aware that their work has broader impact and can be abused (Hovy and Spruit, 2016).

- **Availability and Overexposure** We have found that a small number of papers introduce new datasets or benchmarks (see 5). Creating and curating new datasets as well as benchmarks are often a time-consuming, expensive and long process, where it is oftentimes easier to utilize existing resources to try out new methods (Hovy, 2018). Similarly there is the phenomenon of *topic overexposure*, where there are waves of seemingly 'popular' re-

en		de	sp	zh	fr	it	ru	tr	pl	nl	ar	pt	da	te	sv	no	ja	hi	sk	ko	he	
	0.3	2.1	2.9	-4.2	1.6	0.8	-1.2	1.5	-0.7	-1.7	-0.7	0.0	-2.0	-2.2	-1.2	-1.2	-1.2	-2.2	-0.5	0.5	0.5	en
0.3		2.7	3.7	1.8	1.8	0.8	0.9	0.9	1.9	-0.1	2.9	0.9	-0.1	-0.0	-0.0	-0.0	-0.0	-0.0	1.0	-0.0	1.0	
2.1	2.7		8.7	5.0	7.0	4.2	4.5	4.6	1.7	1.7	2.7	2.7	0.7	-0.2	-0.2	-0.2	1.8	-0.2	-0.1	0.9	1.9	de
2.9	3.7	8.7		3.2	6.2	5.4	3.6	3.7	1.7	-0.3	3.7	3.8	0.8	-0.2	-0.2	-0.2	0.8	-0.2	-0.1	0.9	1.9	sp
-4.2	1.8	5.0	3.2		1.4	0.5	1.7	1.7	0.8	-0.2	2.8	1.8	0.8	-0.1	0.9	-0.1	1.9	-0.1	-0.1	0.9	-0.1	zh
1.6	1.8	7.0	6.2	1.4		6.6	1.7	-0.2	-0.2	0.8	1.8	0.8	0.8	-0.1	-0.1	0.9	-0.1	-0.1	-0.1	-0.1	0.9	fr
0.8	0.8	4.2	5.4	0.5	6.6		1.8	0.8	0.8	0.8	0.8	1.9	0.9	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	0.9	it
-1.2	0.9	4.5	3.6	1.7	1.7	1.8		1.9	-0.1	-0.1	1.9	0.9	0.9	-0.1	0.9	-0.1	0.9	-0.1	-0.0	-0.0	1.0	ru
1.5	0.9	4.6	3.7	1.7	-0.2	0.8	1.9		1.9	-0.1	-0.1	0.9	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.0	1.0	-0.0	tr
-0.7	1.9	1.7	1.7	0.8	-0.2	0.8	-0.1	1.9		-0.1	-0.1	1.9	-0.1	-0.0	-0.0	-0.0	-0.0	-0.0	1.0	-0.0	-0.0	pl
-1.7	-0.1	1.7	-0.3	-0.2	0.8	0.8	-0.1	-0.1	-0.1		-0.1	-0.1	-0.1	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	nl
-0.7	2.9	2.7	3.7	2.8	1.8	0.8	1.9	-0.1	-0.1	-0.1		0.9	-0.1	-0.0	-0.0	-0.0	1.0	-0.0	-0.0	-0.0	1.0	ar
0.0	0.9	2.7	3.8	1.8	0.8	1.9	0.9	0.9	1.9	-0.1	0.9		1.0	-0.0	-0.0	-0.0	1.0	-0.0	-0.0	-0.0	-0.0	pt
-2.0	-0.1	0.7	0.8	0.8	0.8	0.9	0.9	-0.1	-0.1	-0.1	-0.1	1.0		-0.0	1.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	da
-2.2	-0.0	-0.2	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1	-0.0	-0.0	-0.0	-0.0	-0.0		-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	te
-1.2	-0.0	-0.2	-0.2	0.9	-0.1	-0.1	0.9	-0.1	-0.0	-0.0	-0.0	-0.0	1.0	-0.0		-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	sv
-1.2	-0.0	-0.2	-0.2	-0.1	0.9	-0.1	-0.1	-0.1	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0		-0.0	-0.0	-0.0	-0.0	-0.0	no
-1.2	-0.0	1.8	0.8	1.9	-0.1	-0.1	0.9	-0.1	-0.0	-0.0	1.0	1.0	-0.0	-0.0	-0.0	-0.0		-0.0	-0.0	-0.0	-0.0	ja
-2.2	-0.0	-0.2	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0		-0.0	-0.0	-0.0	hi
-0.5	1.0	-0.1	-0.1	-0.1	-0.1	-0.1	-0.0	-0.0	1.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0		-0.0	-0.0	sk
0.5	-0.0	0.9	0.9	0.9	-0.1	-0.1	-0.0	1.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0		-0.0	ko
0.5	1.0	1.9	1.9	-0.1	0.9	0.9	1.0	-0.0	-0.0	-0.0	1.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0		he

Table 4: Residuals between observed joints and predictions based on margins.

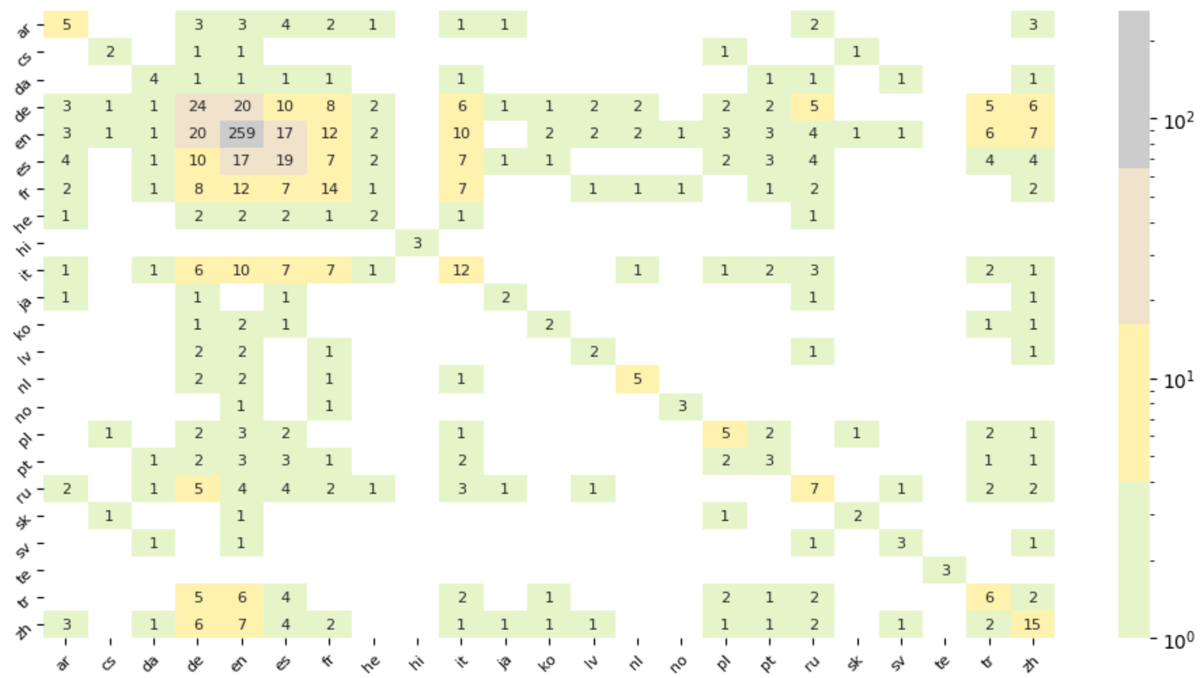

Figure 2: A log-scaled heatmap showing the different languages (ISO-639-1) and their frequency of occurrence.

search topics that eventually go out of fashion. This is based on availability heuristic, if people recall a certain event or have knowledge about certain things then it must be important (Hovy and Spruit, 2016).

Bias We have found a limited focus on specific social biases, where possible causes are rooted in (i) the data that encodes bias by default (Chandrabose et al., 2021), where already available data determines what kind of bias we focus on, (ii) machine learning breakthroughs in NLP has enabled 'streetlamp science' and we focus on tasks that can be solved (Hovy, 2018) and (iii) lack of awareness. This has the consequence that difficult tasks are not being tackled and bias remains present in NLP tools. Therefore it is key to raise awareness (Baeza-Yates, 2018), understand and measure what kind of bias has influence on NLP models and work towards developing solutions that are equitable. Here, we (i) showcase challenges in three frequently researched social biases that have been identified through this survey and (ii) point out opportunities for the future with the aim to raise awareness.

- **Gender bias** There is a strong emphasis on a binary understanding of gender (Schnoebelen, 2017; Orgad and Belinkov, 2022) and most task have been reduced to a masculine/feminine dichotomy (Savoldi et al., 2021). Initially, this may be perceived as useful to enable research, however it does not capture the reality of the society and world we live in today. For example, in the USA alone over 1.4 million people identify as transgender (Larson, 2017) and 1.2 million identify as non-binary (Williams Institute, 2021). At the same time, there is not only a common misconception that gender and sex are the same (Larson, 2017), but also that sexuality is somehow indicative of either gender or sex. Sexuality refers to a person's attraction to a sex or gender, but is not a marker of gender/sex itself (Baum and Westheimer, 2015). Therefore, one can not use sex or gender as a predictor or precursor to assuming a person's sexuality. Talat et al. (2022b) argue that characteristics like sexuality are usually not observable, which can lead to a reliance on hegemonic stereotypes and unnatural language in bias evaluation benchmarks. This leaves plenty of opportunity to start conversations around developing new datasets, benchmarks and methods that are more inclusive and reflect the world we live in (Savoldi et al., 2021; Orgad and Belinkov, 2022).

- **Race bias** Related work by (Field et al., 2021) provides an excellent overview of the state of the art of race bias research in NLP. In their survey they identify that there are very few datasets and benchmarks and that oftentimes a narrow view of race and racial identity are perpetuated. Additionally, researchers often doesn't explicitly state if they are focusing on racial bias through downstream tasks such as abusive language detection. Subsequently, currently deployed hate speech or toxicity classifiers mislabel language predominantly used in the African American community as toxic or hate speech when it is not (Dixon et al., 2018; Xia et al., 2020).

Finally, this survey mentions a number of social biases that have been mentioned such as *religious*, *age* and *disability* with few papers in Figure 1. It is outside of the scope of this work to address each social bias individually, but we would like to point out that there is a lack of relevant benchmarks, datasets and surveys to make substantial progress in these areas and understand the unique challenges each community faces (individually and at an intersectional-level). Most importantly, we would like to emphasize that this type of future work needs to be deeply grounded in interdisciplinary research and led by diverse teams that connect and engage with relevant communities.

Interdisciplinary research The relationship between language and social hierarchies is far more complex than what current techniques can capture. Therefore new methods need to be grounded in relevant literature outside of NLP (Blodgett et al., 2020), because social bias is a complex issue (Sun et al., 2019). Whilst NLP researchers may be committed to using ethical approaches, they may not necessarily have the required ethical and legal knowledge to do so (Santy et al., 2021). This makes it incredibly important to foster collaboration between disciplines to ensure that historical inequalities and biases are taken into consideration when building new algorithms or systems (Caliskan et al., 2017).

Diversity Given the real-life impact of NLP systems and research on people, there is not just a need for diversity in experts working on such systems (Caliskan et al., 2017), but also a need for practitioners and researchers to engage with the affected communities and stakeholders (Blodgett et al., 2020; Fortuna et al., 2021). Therefore, we need to recognize the implicit bias of the people working on different NLP systems and sense-check at different stages how this bias may be reflected in collected data, new benchmarks or models (Savoldi et al., 2021; Hovy and Spruit, 2016). We also need to acknowledge the lack of diversity in teams working on NLP (Schluter, 2018; Savoldi et al., 2021) and work towards more inclusive teams that represent a wide variety of backgrounds and lived experiences (Field et al., 2021). Otherwise, NLP systems continue to represent majorities and we risk the further oppression of already disadvantaged communities (Talat et al., 2022b; Schnoebelen, 2017).

Limitations and Ethics Statement

In this paper, we surveyed a collection of papers and identified continued challenges in social bias research. We have created this collection based on a keyword search and outlined how this may not fully reflect all literature on social bias existing in the ACL anthology or other venues. We only used open-access papers in this collection and no human participants were involved in this work. Traditionally, social biases have been investigated in fields such as social sciences, law, or psychology which we have not discussed here. Furthermore, we do not give an analysis of algorithmic or dataset biases (e.g., machine learning, data mining or otherwise) or provided an in-depth review of technical contributions in computational social biases. We are also limited by the resource of the reviewed papers, where substantial contributions to the field have been made outside of ACL venues. Finally, we would like to point out that opportunities and recommendations for future bias research as proposed in section 6 should be considered from a euro- and/ or anglo-centric perspective. There may be a variation depending on the social context, country or culture that works on a specific bias problem.

Acknowledgements

References

Samyak Agrawal, Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2022. Towards detecting political bias in hindi news articles. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 239–244.

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2932–2949.

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. 2022. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of distilbert. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 266–272.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in bert. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549.

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno Schneider, and Georg Rehm. 2021. Fine-grained classification of political bias in german news: A data set and initial experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131.

Afra Feyza Akyürek, Muhammed Yusuf Kocyigit, Sejin Paik, and Derry Wijaya. 2022a. Challenges in measuring bias via open-ended language generation. *arXiv preprint arXiv:2205.11601*.

Afra Feyza Akyürek, Muhammed Yusuf Kocyigit, Sejin Paik, and Derry Tanti Wijaya. 2022b. Challenges in measuring bias via open-ended language generation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 76–76, Seattle, Washington. Association for Computational Linguistics.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190.

Laura Alonso Alemany, Luciana Benotti, Hernán Maina, Lucía González, Lautaro Martínez, Beatriz Busaniche, Alexia Halvorsen, Amanda Rojo, and Mariela Rajngewerc. 2023. Bias assessment for experts in discrimination, not in computer science. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 91–106.

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830.

Saied Alshahrani, Esma Wali, Abdullah R Alshamsan, Yan Chen, and Jeanna Matthews. 2022. Roadblocks in gender bias measurement for diachronic corpora. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 140–148.

Karl Ameriks and Desmond M Clarke. 2000. *Aristotle: Nicomachean Ethics*. Cambridge University Press.

Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2022a. Sodapop: Open-ended discovery of social biases in social commonsense reasoning models. *arXiv preprint arXiv:2210.07269*.

Haozhe An, Xiaojiang Liu, and Donald Zhang. 2022b. Learning bias-reduced word embeddings using dictionary definitions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1139–1152.

Talita Anthonio and Lennart Kloppenburg. 2019. Team kermit-the-frog at semeval-2019 task 4: Bias detection through sentiment analysis and simple linguistic features. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1016–1020.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119.

Hossein Azarpanah and Mohsen Farhadloo. 2021. Measuring biases of word embeddings: What similarity measures and descriptive statistics to use? In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 8–14.

Ricardo Baeza-Yates. 2018. Bias on the Web. *Communications of the ACM*, 61(6):54–61.

Sunyam Bagga and Andrew Piper. 2020. Measuring the effects of bias in training data for literary classification. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 74–84.

Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen Fraser. 2022. Challenges in applying explainability methods to improve the fairness of NLP models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 80–92, Seattle, U.S.A. Association for Computational Linguistics.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020a. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020b. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3364–3374.

Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*.

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2015. How topic biases your results? a case study of sentiment analysis and irony detection in italian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 41–47.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16.

Christine Basta, Marta R Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39.

Lisa Bauer, Hanna Tischer, and Mohit Bansal. 2023. Social commonsense for explanation and cultural bias discovery. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3727–3742.

Joel Baum and Kim Westheimer. 2015. Sex? Sexual Orientation? Gender Identity? Gender Expression? *Teaching Tolerance*, 50(34–38).

Emily M Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M Bender, Dirk Hovy, and Alexandra Schofield. 2020. Integrating ethics into the nlp curriculum. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–9.

Luciana Benotti and Patrick Blackburn. 2022. Ethics consideration sections in natural language processing papers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4509–4516.

Luciana Benotti, Karën Fort, Min-Yen Kan, and Yulia Tsvetkov. 2023. Understanding ethics in nlp authoring and reviewing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–24.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102.

Amanda Bertsch, Ashley Oh, Sanika Natu, Swetha Gangu, Alan W Black, and Emma Strubell. 2022. Evaluating gender bias transfer from film data. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 235–243.

Advait Bhat, Saaket Agashe, and Anirudha Joshi. 2021. How do people interact with biased text prediction models while writing? In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 116–121.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems*, 29.

Conrad Borchers, Dalia Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Kirk. 2022. Looking for a handsome carpenter! debiasing gpt-3 job advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 212–224.

Shikha Bordia and Samuel Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15.

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A multilingual dataset of racial stereotypes in social media conversational threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 674–684.

Nadine Braun, Martijn Goudbeek, and Emiel Krahmer. 2016. The multilingual affective soccer corpus (masc): Compiling a biased parallel corpus on soccer reportage in english, german and dutch. In *Proceedings of the 9th International Natural Language Generation Conference*, pages 74–78.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1664–1674.

Pere-Lluís Huguet Cabot, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2021. Us vs. them: A dataset of populist attitudes, news bias and emotions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1921–1945.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. 2022. Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in english, spanish, and arabic. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 90–106.

Yang Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022a. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570.

Yang Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022b. Theory-grounded measurement of us social stereotypes in english language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295.

Yang Trista Cao and Hal Daumé III. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle. *Computational Linguistics*, 47(3):615–661.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32.

Aravindan Chandrabose, Bharathi Raja Chakravarthi, et al. 2021. An overview of fairness in data–illuminating the bias in data pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 34–45.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*.

John Chen, Ian Berlot-Attwell, Xindi Wang, Safwan Hossain, and Frank Rudzicz. 2020a. Exploring text specific and blackbox fairness algorithms in multimodal clinical nlp. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 301–312.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2020b. Detecting media bias in news articles using gaussian bias distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4290–4300.

Wei-Fan Chen, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Learning to flip the bias of news headlines. In *Proceedings of the 11th International conference on natural language generation*, pages 79–88.

Xiuying Chen, Mingzhe Li, Rui Yan, Xin Gao, and Xiangliang Zhang. 2022. Unsupervised mitigating gender bias by character components: A case study of chinese word embedding. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 121–128.

Lu Cheng, Ahmadreza Mosallanezhad, Yasin Silva, Deborah Hall, and Huan Liu. 2021. Mitigating bias in session-based cyberbullying detection: A non-compromising approach. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2158–2168.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181.

Jonathan Gabel Christiansen, Mathias Gammelgaard, and Anders Søgaard. 2021. The effect of round-trip translation on fairness in sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4423–4428.

Yung-Sung Chuang, Mingye Gao, Hongyin Luo, James Glass, Hung-Yi Lee, Yun-Nung Chen, and Shang-Wen Li. 2021. Mitigating biases in toxic language detection through invariant rationalization. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 114–120.

Chloe Ciora, Nur Iren, and Malihe Alikhani. 2021. Examining covert gender bias: A case study in turkish and english machine translation models. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 55–63.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32–42.

Kimberlé Williams Crenshaw. 2013. Mapping the margins: Intersectionality, identity politics, and violence against women of color. In *The public nature of private violence*, pages 93–118. Routledge.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Joke Daems and Janiça Hackenbuchner. 2022. Debiasbyus: Raising awareness and creating a database of mt bias. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 287–288.

Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward cultural bias evaluation datasets: The case of bengali gender, religious, and national identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 68–83.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35.

Hillary Dawkins. 2021a. Marked attribute bias in natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4214–4226.

Hillary Dawkins. 2021b. Second order winobias (sowinobias) test set for latent gender bias detection in coreference resolution. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 103–111.

Erenay Dayanık and Sebastian Padó. 2020. Masking actor information leads to fairer political claims detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4385–4391.

FMG De Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, Dieter Van Uytvanck, et al. 2018. Clarin: towards fair and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3259–3264.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2021a. Oscar: Orthogonal subspace correction and rectification of biases in word embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5034–5050.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2021b. What do bias measures measure?

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020. Semi-supervised topic modeling for gender bias discovery in english and swedish. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020b. Multidimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Jad Doughman, Wael Khreich, Maya El Gharib, Maha Wiss, and Zahraa Berjawi. 2021. Gender bias in text: Origin, taxonomy, and implications. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 34–44.

Jo Drugan and Bogdan Babych. 2010. Shared resources, shared values? ethical implications of sharing translation resources. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 3–10.

Yupei Du, Qixiang Fang, and Dong Nguyen. 2021. Assessing the reliability of word embedding gender bias measures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10012–10034.

Yupei Du, Qi Zheng, Yuanbin Wu, Man Lan, Yan Yang, and Meirong Ma. 2022. Understanding gender bias in knowledge base embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1381–1395.

Penelope Eckert and John R Rickford. 2001. *Style and sociolinguistic variation*. Cambridge University Press.

Fatma Elsafoury, Steven R Wilson, and Naeem Ramzan. 2022. A comparative study on word embeddings and social nlp tasks. In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, pages 55–64.

Carla Parra Escartın, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. Ethical considerations in nlp shared tasks. *EACL 2017*, page 66.

Kawin Ethayarajh. 2020. Is your classifier actually biased? measuring fairness under uncertainty with bernstein bounds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2914–2919.

Agnieszka Falenska and Özlem Çetinoğlu. 2021. Assessing gender bias in Wikipedia: Inequalities in article titles. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 75–85, Online. Association for Computational Linguistics.

Angela Fan and Claire Gardent. 2022. Generating biographies on wikipedia: The impact of gender bias on the retrieval-based generation of women biographies. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8561–8576.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349.

Michael Färber, Agon Qurdina, and Lule Ahmedi. 2019. Team peter brinkmann at semeval-2019 task 4: Detecting biased news articles using convolutional neural networks. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1032–1036.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608.

Joseph Fisher, Arpit Mittal, Dave Palfrey, and Christos Christodoulopoulos. 2020. Debiasing knowledge graph embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7332–7345.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854.

Joel Escudé Font and Marta R Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154.

Karën Fort and Alain Couillault. 2016. Yes, we care! results of the ethics and natural language processing surveys. In *international Language Resources and Evaluation Conference (LREC) 2016*.

Paula Fortuna, Laura Pérez-Mayos, Leo Wanner, et al. 2021. Cartography of natural language processing for social good (nlp4sg): Searching for definitions, statistics and white spots. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 19–26.

Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. 2019. Relating word embedding gender biases to gender gaps: A cross-cultural analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 18–24.

Annemarie Friedrich and Torsten Zesch. 2021. A crash course on ethics for natural language processing. In *Proceedings of the Fifth Workshop on Teaching NLP*, pages 49–51.

Niklas Friedrich, Anne Lauscher, Simone Paolo Ponzetto, and Goran Glavaš. 2021. Debie: A platform for implicit and explicit debiasing of word embedding spaces. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 91–98.

Zee Fryer, Vera Axelrod, Ben Packer, Alex Beutel, Jilin Chen, and Kellie Webster. 2022. Flexible text generation for counterfactual fairness probing. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 209–229.

Yacine Gaci, Boualem Benattallah, Fabio Casati, and Khalid Benabdeslem. 2022. Debiasing pretrained text encoders by paying attention to paying attention. In *2022 Conference on Empirical Methods in Natural Language Processing*, pages 9582–9602. Association for Computational Linguistics.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. How to split: the effect of word segmentation on gender bias in speech translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3576–3589.

Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, and Mohit Iyyer. 2020. Analyzing gender bias within narrative tropes. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 212–217.

Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. 2019. Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pages 77–84.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, N Anandhavelu, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep NLP. *Applied Sciences*, 11(7):3184.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, et al. 2020. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328.

Noa Baker Gillis. 2021. Sexism in the judiciary: The importance of bias definition in nlp and in our courts. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 45–54.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type b reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57.

Meiqi Guo, Rebecca Hwa, Yu-Ru Lin, and Wen-Ting Chung. 2020. Inflating topic relevance with ideology: A case study of political ideology bias in social topic detection models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4873–4885.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023.

Saurabh Gupta, Hong Huy Nguyen, Junichi Yamagishi, and Isao Echizen. 2020. Viable threat on news reading: Generating biased news using natural language models. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 55–65.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 658–678.

Gerhard Hagerer, David Szabo, Andreas Koch, Maria Luisa Ripoll Dominguez, Christian Widmer, Maximilian Wich, Hannah Danner, and Georg Groh. 2021. End-to-end annotator bias approximation on crowdsourced single-label sentiment analysis. In *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 1–10.

Felix Hamborg. 2020. Media bias, the social sciences, and nlp: automating frame analyses to identify bias by word choice and labeling. In *Proceedings of the 58th annual meeting of the association for computational linguistics: student research workshop*, pages 79–87.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Decoupling adversarial training for fair nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 471–477.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2023. Fair enough: Standardizing evaluation and model selection for fairness research in nlp. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 297–312.

Oussama Hansal, Ngoc Tan Le, and Fatiha Sadat. 2022. Indigenous language revitalization and the dilemma of gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 244–254.

Victor Petrén Bach Hansen and Anders Søgaard. 2021. Guideline bias in wizard-of-oz dialogues. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 8–14.

Levon Haroutunian. 2022. Ethical considerations for low-resourced machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 44–54.

Lucy Havens, Beatrice Alex, Benjamin Bach, and Melissa Terras. 2022. Uncertainty and inclusivity in gender bias annotation: An annotation taxonomy and annotated datasets of british english text. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 30–57.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. Mabel: Attenuating gender bias using textual entailment data. *arXiv preprint arXiv:2210.14975*.

Zexue He, Bodhisattwa Prasad Majumder, and Julian McAuley. 2021. Detect and perturb: Neutral rewriting of biased and sensitive text via gradient-based decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4173–4181.

Brienna Herold, James Waller, and Raja Kushalnagar. 2022. Applying the stereotype content model to assess disability bias in popular pre-trained nlp models underlying ai-based assistive technologies. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 58–65.

Livnat Herzig, Alex Nunes, and Batia Snir. 2011. An annotation scheme for automated bias detection in wikipedia. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 47–55.

Marius Hessenthaler, Emma Strubell, Dirk Hovy, and Anne Lauscher. 2022. Bridging fairness and environmental sustainability in natural language processing. *arXiv preprint arXiv:2211.04256*.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 752–762.

Dirk Hovy. 2018. The social and the neural network: How to make natural language processing about people again. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 42–49.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020a. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering implicit gender bias in narratives through commonsense inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3866–3873.

Xiaolei Huang. 2022. Easy adaptation to mitigate gender bias in multilingual text classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 717–723.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael Paul. 2020b. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1440–1448.

Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 49–58.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Sepehr Janghorbani and Gerard De Melo. 2023. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1717–1727.

Sophie Jentzsch and Cigdem Turan. 2022. Gender bias in bert-measuring and analysing biases through sentiment rating in a realistic downstream classification task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199.

Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. 2020. Mitigating gender bias amplification in distribution by posterior regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2936–2942.

Meichun Jiao and Ziyang Luo. 2021. Gender bias hidden behind chinese word embeddings: The case of chinese adjectives. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 8–15.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021a. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021b. How good is NLP? A sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113.

Przemyslaw Joniak and Akiko Aizawa. 2022. Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 67–73.

Lalitha Kameswari and Radhika Mamidi. 2021. Towards quantifying magnitude of political bias in news articles using a novel annotation schema. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 671–678.

Lalitha Kameswari, Dama Sravani, and Radhika Mamidi. 2020. Enhancing bias detection in political news using pragmatic presupposition. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 1–6.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650.

Masahiro Kaneko and Danushka Bollegala. 2021a. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266.

Masahiro Kaneko and Danushka Bollegala. 2021b. Dictionary-based debiasing of pre-trained word embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 212–223.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2023. Comparing intrinsic gender bias evaluation measures without using human annotated examples. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2849–2855.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on weat. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48.

Daphna Keidar, Mian Zhong, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. 2021. Towards automatic bias detection in knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3804–3811.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22.

Satyapriya Krishna, Rahul Gupta, Apurv Verma, Jwala Dhamala, Yada Pruksachatkun, and Kai-Wei Chang. 2022. Measuring fairness of text classifiers via prediction sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5830–5842.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023. Parameter-efficient modularised bias mitigation via adapterfusion. In *Proceedings of the 17th Conference of the European Chapter of* *the Association for Computational Linguistics*, pages 2730–2743.

Vaibhav Kumar and Tenzin Bhotia. 2020. Fair embedding engine: A library for analyzing and mitigating gender bias in word embeddings. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 26–31.

Vaibhav Kumar, Tenzin Singhay Bhotia, and Tanmoy Chakraborty. 2020. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *Transactions of the Association for Computational Linguistics*, 8:486–503.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Alexander Kwako, Yixin Wan, Jieyu Zhao, Kai-Wei Chang, Li Cai, and Mark Hansen. 2022. Using item response theory to measure gender and racial bias of a bert-based automated english speech assessment system. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 1–7.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen Mckeown, and Tatsunori B Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3198–3211.

John P Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609.

Julian Lamont and Christi Favor. 2004. Distributive justice. *Handbook of political theory*, 1.

Brian N Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *First Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. *NAACL HLT 2019*, page 85.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797.

Konstantina Lazaridou, Alexander Löser, Maria Mestre, and Felix Naumann. 2020. Discovering biased news articles leveraging multiple human annotations. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1268–1277.

Jochen L Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40.

Heather Lent and Anders Søgaard. 2021. Common sense bias in semantic role labeling. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 114–119.

Michael Lepori. 2020. Unequal representations: Analyzing intersectional biases in word embeddings using representational similarity analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1720–1728.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480.

Dave Lewis, Joss Moorkens, and Kaniz Fatema. 2017. Integrating the management of personal data protection and open science with research ethics. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 60–65.

Jiali Li, Shucheng Zhu, Ying Liu, and Pengyuan Liu. 2022a. Analysis of gender bias in social perception and judgement using chinese word embeddings. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 8–16.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. Unqovering stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489.

Yuantong Li, Xiaokai Wei, Zijian Wang, Shen Wang, Parminder Bhatia, Xiaofei Ma, and Andrew Arnold. 2022b. Debiasing neural retrieval via in-batch balancing regularization. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 58–66.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1478–1484.

Tomasz Limisiewicz and David Mareček. 2022. Don't forget about pronouns: Removing gender bias in language models without losing factual gender information. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 17–29.

Emmy Liu, Michael Henry Tessler, Nicole Dubosh, Katherine Hiller, and Roger Levy. 2022a. Assessing group-level gender bias in professional evaluations: The case of medical student end-of-shift feedback. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 86–93.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020a. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416.

Haochen Liu, Wei Jin, Hamid Karimi, Zitao Liu, and Jiliang Tang. 2021. The authors matter: Understanding and mitigating implicit bias in deep text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 74–85.

Haochen Liu, Joseph Thekinen, Sinem Mollaoglu, Da Tang, Ji Yang, Youlong Cheng, Hui Liu, and Jiliang Tang. 2022b. Toward annotator group bias in crowdsourcing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1797–1806.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020b. Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903.

Yunfei Long, Mingyu Ma, Qin Lu, Rong Xiang, and Chu-Ren Huang. 2018. Dual memory network model for biased product review classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 140–148.

Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10.

Li Lucy and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55.

Hongyin Luo and James Glass. 2023. Logic against bias: Textual entailment mitigates stereotypical sentence reasoning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1235–1246.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. Powertransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441.

Nitin Madnani, Anastassia Loukina, Alina Von Davier, Jill Burstein, and Aoife Cahill. 2017. Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially aware bias measurements for hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052.

Courtney Mansfield, Amandalynne Paullada, and Kristen Howell. 2022. Behind the mask: Demographic bias in name detection for pii masking. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 76–89.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of NAACL-HLT*, pages 615–621.

Andrew Mao, Naveen Raman, Matthew Shu, Eric Li, Franklin Yang, and Jordan Boyd-Graber. 2021. Eliciting bias in question answering models through ambiguity. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 92–99.

Sanjana Marcé and Adam Poliak. 2022. On gender biases in offensive language classification models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 174–183.

Sandra Martinková, Karolina Stanczak, and Isabelle Augenstein. 2023. Measuring gender bias in west slavic language models. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 146–154.

Judith Masthoff. 2011. Book review: Close engagements with artificial companions: Key social, psychological, ethical, and design issues edited by yorick wilks. *Computational Linguistics*, 37(2).

Camila M Mateo and David R Williams. 2020. Addressing bias and reducing discrimination: The professional responsibility of health care providers. *Academic Medicine*, 95(12S):S5–S10.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628.

Elijah Mayfield, Michael Madaio, Shrimai Prabhumoye, David Gerritsen, Brittany McLaughlin, Ezekiel Dixon-Román, and Alan W Black. 2019. Equity beyond bias in language technologies for education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 444–460.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898.

Michal Měchura. 2022. A taxonomy of bias-causing ambiguities in machine translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 168–173.

Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. 2022. Attributing fair decisions with attention interventions. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 12–25.

Jack Merullo, Luke Yeh, Abram Handler, Alvin Grissom II, Brendan O'Connor, and Mohit Iyyer. 2019. Investigating sports commentator bias within a large corpus of american football broadcasts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6355–6361.

Josh Meyer, Lindy Rauchenstein, Joshua D Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6462–6468.

Piotr Miłkowski, Marcin Gruza, Kamil Kanclerz, Przemyslaw Kazienko, Damian Grimling, and Jan Kocoń. 2021. Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 248–259.

Fatemehsadat Mireshghallah and Taylor Berg-Kirkpatrick. 2021. Style pooling: Automatic text style obfuscation for improved classification fairness. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2009–2022.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2021. Modeling users and online communities for abuse detection: A position on ethics and explainability. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3374–3385.

Saif Mohammad. 2022a. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Saif Mohammad. 2022b. Ethics sheets for AI tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8368–8379.

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in dutch word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 56–71.

Robert Munro and Alex Carmen Morrison. 2020. Detecting independent pronoun bias with partially-synthetic data generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2011–2017.

Taichi Murayama, Shoko Wakamiya, and Eiji Aramaki. 2021. Mitigation of diachronic bias in fake news detection dataset. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 182–188.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French crows-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Timothy Niven and Hung-Yu Kao. 2020. Measuring alignment to authoritarian state media as framing bias. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 11–21.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 68–74.

Hadas Orgad and Yonatan Belinkov. 2022. Choose your lenses: Flaws in gender bias evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167, Seattle, Washington. Association for Computational Linguistics.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628.

Martin Orr, Kirsten Van Kessel, and Dave Parry. 2022. The ethical role of computational linguistics in digital psychological formulation and suicide prevention. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 17–29.

Prasanna Parasurama and João Sedoc. 2022. Gendered language in resumes and its implications for algorithmic bias in hiring. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 74–74.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2022. Don't blame the annotator: Bias already starts in the annotation instructions. *arXiv preprint arXiv:2205.00415*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.

Jannik Pedersen, Martin Laursen, Pernille Vinholt, Anne Alnor, and Thiusius Savarimuthu. 2023. Investigating anatomical bias in clinical machine learning algorithms. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1368–1380.

Matúš Pikuliak, Ivana Beňová, and Viktor Bachratỳ. 2023. In-depth look at word filling societal bias measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3630–3647.

Shrimai Prabhumoye, Brendon Boldt, Ruslan Salakhutdinov, and Alan W Black. 2021. Case study: Deontological ethics in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3784–3798.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75.

Ming Qian, Jessie Liu, Chaofeng Li, and Liming Pals. 2019a. A comparative study of english-chinese translations of court texts by machine and human translators and the word2vec based similarity measure's ability to gauge human evaluation biases. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 95–100.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019b. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228.

Krithika Ramesh, Gauri Gupta, and Sanjay Singh. 2021. Evaluating gender bias in hindi-english machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond english: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2061–2074.

Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040. Association for Computational Linguistics.

Adithya Renduchintala, Denise Díaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109.

Alisa Rieger, Mariët Theune, and Nava Tintarev. 2020. Toward natural language mitigation strategies for cognitive biases in recommender systems. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, pages 50–54.

Anthony Rios, Reenam Joshi, and Hejin Shin. 2020. Quantifying 60 years of gender bias in biomedical research with word embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? reducing bias in bios without access to protected attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4187–4195.

Candace Ross, Boris Katz, and Andrei Barbu. 2021. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of NAACL-HLT*, pages 8–14.

Magnus Sahlgren and Fredrik Olsson. 2019. Gender bias in pretrained swedish embeddings. In *Proceedings of the 22nd Nordic Conference on computational linguistics*, pages 35–43.

Sebastin Santy, Anku Rani, and Monojit Choudhury. 2021. Use of formal ethical reviews in NLP literature: Historical trends and current practices. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4704–4710.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824.

Ramit Sawhney, Arshiya Aggarwal, and Rajiv Shah. 2021. An empirical investigation of bias in the multimodal analysis of financial earnings calls. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3751–3757.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Natalie Schluter. 2018. The glass ceiling in NLP. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2793–2798.

Katja Geertruida Schmahl, Tom Julian Viering, Stavros Makrodimitris, Arman Naseri Jahfari, David Tax, and Marco Loog. 2020. Is wikipedia succeeding in reducing gender bias? assessing changes in gender bias in wikipedia using word embeddings. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 94–103.

Tyler Schnoebelen. 2017. Goal-oriented design for ethical machine learning and NLP. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 88–93.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. "This is a problem, don't you agree?" Framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16.

Lena Schwertmann, Manoj Prabhakar Kannan Ravi, and Gerard De Melo. 2023. Model-agnostic bias measurement in link prediction. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1587–1603.

João Sedoc and Lyle Ungar. 2019. The role of protected class word lists in bias identification of contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 55–61.

Indira Sen, Mattia Samory, Fabian Flöck, Claudia Wagner, and Isabelle Augenstein. 2021. How does counterfactually augmented data impact models for social computing constructs? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 325–344.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4716–4726.

Emeralda Sesari, Max Hort, and Federica Sarro. 2022. An empirical study on the fairness of pre-trained word embeddings. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 129–144.

Usman Shahid, Barbara Di Eugenio, Andrew Rojecki, and Elena Zheleva. 2020. Detecting and understanding moral biases in news. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 120–125.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Optimising equal opportunity fairness in model training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4073–4084.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293.

Emily Sheng and David C Uthus. 2020. Investigating societal biases in a poetry composition system. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 93–106.

Seungjae Shin, Kyungwoo Song, JoonHo Jang, Hyemi Kim, Weonyoung Joo, and Il-Chul Moon. 2020. Neutralizing gender bias in word embeddings with latent disentanglement and counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3126–3140.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389.

Charese Smiley, Frank Schilder, Vassilis Plachouras, and Jochen L Leidner. 2017. Say the right thing right: Ethics issues in natural language generation systems. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 103–108.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with babe-bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87.

Tejas Srinivasan and Yonatan Bisk. 2022a. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85, Seattle, Washington. Association for Computational Linguistics.

Tejas Srinivasan and Yonatan Bisk. 2022b. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542.

Angus Stevenson. 2010. *Oxford dictionary of English*. Oxford University Press, USA.

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021a. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2498.

Shivashankar Subramanian, Afshin Rahimi, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021b. Fairness-aware class imbalanced learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2045–2051.

Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. Bertscore is unfair: On social bias in language model-based metrics for text generation. *arXiv preprint arXiv:2210.07626*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 80–87, Valencia, Spain. Association for Computational Linguistics.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667.

Masashi Takeshita, Yuki Katsumata, Rafal Rzepka, and Kenji Araki. 2020. Can existing methods debias languages other than english? first attempt to analyze and mitigate japanese word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 44–55.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022a. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779.

Zeerak Talat, Aurelie Neveol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, et al. 2022b. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41.

Rachael Tatman. 2017a. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59.

Rachael Tatman. 2017b. "oh, i've heard that before": Modelling own-dialect bias after perceptual learning by weighting training data. In *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017)*, pages 29–34.

Wenyi Tay. 2019. Not all reviews are equal: Towards addressing reviewer biases for opinion summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 34–42.

Andamlak Terkik, Emily Prud'hommeaux, Cecilia Ovesdotter Alm, Christopher Homan, and Scott Franklin. 2016. Analyzing gender bias in student evaluations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 868–876.

Ewoenam Tokpo, Pieter Delobelle, Bettina Berendt, and Toon Calders. 2023. How far can it go? on intrinsic gender bias mitigation for text classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3410–3425.

Ewoenam Kwaku Tokpo and Toon Calders. 2022. Text style transfer for bias mitigation using masked language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 163–171.

Autumn Toney and Aylin Caliskan. 2021. Valnorm quantifies semantics to reveal consistent valence biases across languages and over centuries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7203–7218.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2021. Using gender-and polarity-informed models to investigate bias. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 66–74.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2022. Occupational biases in norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2023. Measuring normative and descriptive biases in language models using census data. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2234–2240.

Jonas Troles and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation impact of stereotypical verbs and adjectives. *WMT 2021*, page 531.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599.

Oskar Van Der Wal, Jaap Jumelet, Katrin Schulz, and Willem Zuidema. 2022. The birth of bias: A case study on the evolution of gender bias in an english language model. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 75–75.

Lindsey Vanderlyn, Gianna Weber, Michael Neumann, Dirk Väth, Sarina Meyer, and Ngoc Thang Vu. 2021. "it seemed like an annoying woman": On the perception and ethical considerations of affective language in text-based conversational agents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 44–57.

Francisco Vargas and Ryan Cotterell. 2020. Exploring the linear subspace hypothesis in gender bias mitigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2902–2913.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122.

Xuan-Son Vu, Thanh-Son Nguyen, Duc-Trong Le, and Lili Jiang. 2020. Multimodal review generation with privacy and fairness awareness. In *28th International Conference on Computational Linguistics (COLING), Barcelona, Spain (Online), December 8-13, 2020.*, pages 414–425. International Committee on Computational LinguisticsInternational Committee ....

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, and Yi Chang. 2021a. Eliminating sentiment bias for aspect-level sentiment classification with unsupervised opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3002–3012.

Jialu Wang, Yang Liu, and Xin Wang. 2021b. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1995–2008.

Jianan Wang, Xin Wang, Fang Li, Zhen Xu, Zhuoran Wang, and Baoxun Wang. 2017. Group linguistic bias aware neural response generation. In *Proceedings of the 9th SIGHAN Workshop on Chinese Language Processing*, pages 1–10.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590.

Liwen Wang, Yuanmeng Yan, Keqing He, Yanan Wu, and Weiran Xu. 2021c. Dynamically disentangling social bias from task-oriented representations with adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3740–3750.

Shuo Wang, Zhaopeng Tu, Zhixing Tan, Shuming Shi, Maosong Sun, and Yang Liu. 2021d. On the language coverage bias for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4778–4790.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming

Xiong. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5443–5453.

Zhao Wang, Kai Shu, and Aron Culotta. 2021e. Enhancing model robustness and fairness with causality: A regularization approach. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 33–43.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in NLP. *arXiv preprint arXiv:2101.11974*.

Melvin Wevers. 2019. Using word embeddings to examine gender bias in dutch newspapers, 1950-1990. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 92–97.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020a. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199.

Maximilian Wich, Jan Bauer, and Georg Groh. 2020b. Impact of politically biased data on hate speech classification. In *Proceedings of the fourth workshop on online abuse and harms*, pages 54–64.

Maximilian Wich, Christian Widmer, Gerhard Hagerer, and Georg Groh. 2021. Investigating annotator bias in abusive language datasets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1515–1525.

Williams Institute. 2021. 1.2 million LGBTQ adults in the US identify as nonbinary.

Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14.

Zhongbin Xie, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2023. Counter-gap: Counterfactual bias evaluation through gendered ambiguous pronouns. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3743–3755.

Tae Yano, Philip Resnik, and Noah A Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 152–158.

Samira Zad, Joshuan Jimenez, and Mark Finlayson. 2021. Hell hath no fury? correcting bias in the nrc emotion lexicon. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 102–113.

Haiyang Zhang, Alison Sneyd, and Mark Stevenson. 2020. Robustness and reliability of gender bias assessment in word embeddings: The role of base pairs. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 759–769.

Jieyu Zhao and Kai-Wei Chang. 2020. Logan: Local group bias detection by clustering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1968–1977.

Jieyu Zhao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Kai-Wei Chang. 2021. Ethical-advice taker: Do language models understand natural language interventions? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4158–4164.

Jieyu Zhao, Subhabrata Mukherjee, Kai-Wei Chang, Ahmed Hassan Awadallah, et al. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Ruiqi Zhong, Yanda Chen, Desmond Patton, Charlotte Selous, and Kathleen Mckeown. 2019. Detecting and reducing bias in a high stakes domain. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4765–4775.

Yang Zhong, Jingfeng Yang, Wei Xu, and Diyi Yang. 2021. Wikibias: Detecting multi-span subjective biases in language. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1799–1814.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang.

2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155.

Yi Zhou, Masahiro Kaneko, and Danushka Bollegala. 2022. Sense embeddings are also biased–evaluating social biases in static and contextualised sense embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1924–1935.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773.

Nadezhda Zueva, Madina Kabirova, and Pavel Kalaidin. 2020. Reducing unintended identity bias in russian hate speech detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 65–69.

7 Appendix

Type	**Papers**
Survey/Position Paper	Drugan and Babych (2010), Masthoff (2011), Escartın et al. (2 (2016), Hovy and Spruit (2016), Leidner and Plachouras (2017), et al. (2017), Schnoebelen (2017), Larson (2017), De Jong et a (2018), Mayfield et al. (2019),Sun et al. (2019), Blodgett et al. (2021), Bender et al. (2020), Schoch et al. (2020), Savoldi et a (2021), Mishra et al. (2021), Czarnowska et al. (2021), Dough Sheng et al. (2021), Hovy and Yang (2021), Santy et al. (2021), (2022b), Chandrabose et al. (2021), Vecchi et al. (2021), Talat et a Balkir et al. (2022), Orgad and Belinkov (2022), Akyürek et al (2022), Mohammad (2022a), Orr et al. (2022), Benotti and Bla et al. (2023), Benotti et al. (2023), Ramesh et al. (2023), Rios et
Dataset, Benchmark	Yano et al. (2010), Flekova et al. (2016), Braun et al. (2016) Ka Rudinger et al. (2018), Zhao et al. (2018), Kiritchenko and Mo (2019) , Gaut et al. (2020), Friedman et al. (2019), Hitti et al. ((2019), Fan et al. (2019), Liu et al. (2020a), Sap et al. (2020), Lou et al. (2021), Lim et al. (2020), Shahid et al. (2020), Lazarido Uthus (2020), Gala et al. (2020), Dinan et al. (2020b), Meyer e (2020b), Gangula et al. (2019), Zueva et al. (2020), Nangia et al. Blodgett et al. (2021), Spinde et al. (2021), Zhong et al. (2021), Søgaard (2021), Zhao et al. (2021), Aksenov et al. (2021), Kam Renduchintala et al. (2021), Barikeri et al. (2021), Miłkowski et et al. (2022), Ziems et al. (2022), Zhou et al. (2022), Akyürek e (2022), Bansal et al. (2022), Smith et al. (2022), Qian et al. (2022 et al. (2023), Madnani et al. (2017), Kumar and Bhotia (2020) (2021), Zad et al. (2021), Nozza et al. (2022), Měchura (2022)
Methods	Gonen and Goldberg (2019), Wevers (2019), Hovy et al. (2020 Nissim et al. (2020), Wang et al. (2021d), Wolfe and Caliskan (2 et al. (2021), Herold et al. (2022) , Bertsch et al. (2022), Ramp Cao et al. (2022a), Alshahrani et al. (2022) Daems and Hackenbu et al. (2015) , Terkik et al. (2016), Tatman (2017b), Wang et al Tatman (2017a), Zhao et al. (2017), Long et al. (2018), Baly e Kloppenburg (2019), Färber et al. (2019), Zhong et al. (2019), T and Ungar (2019), Gangula et al. (2019), Karve et al. (2019), I and Najafian (2019), Chaloner and Maldonado (2019) , Qian et (2019) , Kaneko and Bollegala (2019), Huang et al. (2020a), Ba Dinan et al. (2020a), Bordia and Bowman (2019) , Manzini et al. , Schwertmann et al. (2023), Pedersen et al. (2023), Bauer et al. (Kumar et al. (2023) , Touileb et al. (2023), Parmar et al. (2022), et al. (2023), Alemany et al. (2023), Martinková et al. (2023), I (2022), Cao et al. (2022b), Sap et al. (2020) , Xia et al. (2020) (2022) , Orgad et al. (2022), Sen et al. (2022), Gupta et al. (202 Gira et al. (2022), Guo et al. (2022), Wang et al. (2022), An et a (2022), Borchers et al. (2022), Touileb et al. (2022), Jentzsch and et al. (2022), Tal et al. (2022), Li et al. (2022b) , Limisiewicz an et al. (2022) , Ahn et al. (2022), Joniak and Aizawa (2022), Chen (2022), Du et al. (2022) , Agrawal et al. (2022), Sesari et al. (202 Steed et al. (2022), Schick et al. (2021), Zhou et al. (2021), Kane et al. (2021) , Lucy and Bamman (2021), Jin et al. (2021a), L (2021b), Silva et al. (2021), Sun and Peng (2021), Wang et al. (2 , Han et al. (2021), Sen et al. (2021), Ramesh et al. (2021), Sub and Oh (2021) , Davidson et al. (2019), May et al. (2019), Sap Piper (2020), Guo et al. (2020) , Vu et al. (2020), Spliethöver a Munro and Morrison (2020), Zhao and Chang (2020), Schmahl (2020), Chen et al. (2020b) , Mulsa and Spanakis (2020), Shir Cotterell (2020), Liang et al. (2020), Wich et al. (2020a) , Jia et a (2020) , Dev et al. (2021a), Kumar et al. (2020), Field and Tsvetk (2020a), Stafanovičs et al. (2020) , Li et al. (2020), Ma et al. (2 Al Kuwatly et al. (2020), Bartl et al. (2020) , Liu et al. (2020b), H , Sheng et al. (2020), Gaci et al. (2022), Hutchinson et al. (2020) et al. (2021e) ,Van Der Wal et al. (2022), Hansen and Søgaard (2 Azarpanah and Farhadloo (2021), Jiao and Luo (2021) , Touileb , Ciora et al. (2021), Murayama et al. (2021), Gaido et al. (2021 (2022), Dawkins (2021a) , Gillis (2021), Mireshghallah and B Wang et al. (2021a), Malik et al. (2022), Subramanian et al. (202 Garimella et al. (2021) , Ghosh et al. (2021), Mehrabi et al. (202
Both above	Goldfarb-Tarrant et al. (2021), Janghorbani and De Melo (202 Troles and Schmid (2021), Liu et al. (2022a), Névéol et al. (202
Analysis	Herzig et al. (2011), Rudinger et al. (2017), Lauscher and Glava Vanderlyn et al. (2021) , Falenska and Çetinoğlu (2021), Sap et a (2021), Bhat et al. (2021), Haroutunian (2022) , Qian et al. (201